Mol Gen Genet (1988) 213:84–92

MGG

Sequence structure and expression of a cloned β-glucosidase gene from an extreme thermophile

D.R. Love, R. Fisher, and P.L. Bergquist

Department of Cell Biology, University of Auckland, Private Bag, Auckland, New Zealand

Summary. The gene for a β-glucosidase from the extremely thermophilic bacterium *Caldocellum saccharolyticum* has been isolated from a genomic library and sequenced. An open reading frame identified by computer analysis of the sequence could encode a protein of M_r 54400, which is close to the size of the polypeptide experimentally determined using maxicells. Analysis of the amino-terminal residues of the protein produced in *Escherichia coli* suggests that it is processed by a methionine aminopeptidase. A sequence within *C. saccharolyticum* DNA upstream of the β-glucosidase gene was found to act as a promoter for expression of the thermophile gene in *E. coli*. The protein has been overproduced in *E. coli* and *Bacillus subtilis* where it retains its enzymatic activity and heat stability. There appears to be a single copy of the gene in *Caldocellum* DNA.

Key words: Thermophile – β-Glucosidase – Sequence analysis – Expression vectors

Introduction

Three general classes of enzymes are involved in the breakdown of cellulose: exocellulase (β-1,4-D-glucan cellobiohydrolase), endocellulase (β-1,4-D-glucan glucanohydrolase) and β-1,4-D-glucosidase. The first two enzymes act co-operatively to depolymerize cellulose to cellobiose and oligosaccharides. β-Glucosidase hydrolyses these sugars to form glucose.

The obligatory anaerobe, *Caldocellum saccharolyticum*, is a thermophilic bacterium that has an optimum growth temperature of 68° C but which will continue to grow at 80° C under laboratory conditions. It is able to degrade cellulose but is unrelated to the intensively studied species *Clostridium thermocellum*, as shown by a lack of DNA-DNA hybridization (Donnison et al. 1986). We have constructed a gene bank of *C. saccharolyticum* in bacteriophage λ1059 and have isolated recombinants that carry DNA encoding a number of enzymes involved in cellulose breakdown. A β-glucosidase from *Caldocellum* has been expressed in *Escherichia coli* and *Bacillus subtilis* (Love and Streiff 1987). The enzyme purified from the mesophilic host has a temperature maximum of 85° C, a pH maximum of 6.25, and M_r 52000, properties identical to those of the enzyme isolated from *Caldocellum*. In this communication, we report the nucleotide sequence of the β-glucosidase gene, the location of a thermophile DNA sequence recognised as a promoter in *E. coli*, and the high level expression of β-glucosidase in *E. coli* and *B. subtilis* hosts.

Offprint requests to: P.L. Bergquist

Materials and methods

Bacteria and plasmids. *E. coli* strains used were: PB2636 (F^-, *galK, thi-1, leu-6, thr-1, lacY1, supE44,* r_k^- m_k^+); JM101 (F′ *traD36,* $proA^+B^+$, *lacZ*ΔM15/Δ*lac pro, thi, supE44*); JM105 (F′ *traD36,* $proA^+B^+$, $lacI^q$, *lacZ*ΔM15/Δ*lac pro, thi, strA, endA, sbcB15, hspR4*). The *B. subtilis* strain used was SB202 (*aroB2, trpC2, tyrA1, hisH2*), provided by D. Ehrlich. The plasmids used were pPL608, provided by P. Lovett, pKK223-3 which was purchased from P-L Biochemicals, and pK0100 (McKenney et al. 1981).

Media, transformation and DNA techniques. These were as described previously (Love and Streiff 1987).

β-Glucosidase assay. Samples of cultures were diluted in 50 mM phosphate-citrate buffer, pH 6.25 and treated with toluene. Appropriate volumes were assayed for β-glucosidase activity at 70° C for 60 min in the presence of 0.5 mg ml^{-1} *p*-nitrophenyl-β-D-glucopyranoside (PNPG, Sigma), as described previously (Love and Streiff 1987). One unit of β-glucosidase activity was defined as the amount of enzyme required to liberate 1 μmol *p*-nitrophenol per minute.

Southern hybridization. DNAs were electrophoresed in a 0.8% agarose-borate gel and transferred to Genescreen Plus (NEN Research Products) following the method described by Reed and Mann (1985). The 1.72 kb *Hin*dIII fragment of pNZ1002 (Fig. 2) was isolated from a gel using Geneclean (Bio 101) and the DNA was nick-translated using a kit purchased from Bethesda Research Laboratories.

SDS-gel electrophoresis. Samples (100 μl) of *B. subtilis* cultures were sonicated for 15 s in the presence of SDS-loading buffer and then boiled for 5 min. Undissolved material was removed by centrifugation and 20 μl samples of the supernatants were electrophoresed in an SDS-polyacrylamide gel according to Laemmli (1970). Proteins were fixed in 50% methanol, 10% acetic acid and visualized by the silver staining method essentially as described by Oakley et al. (1980).

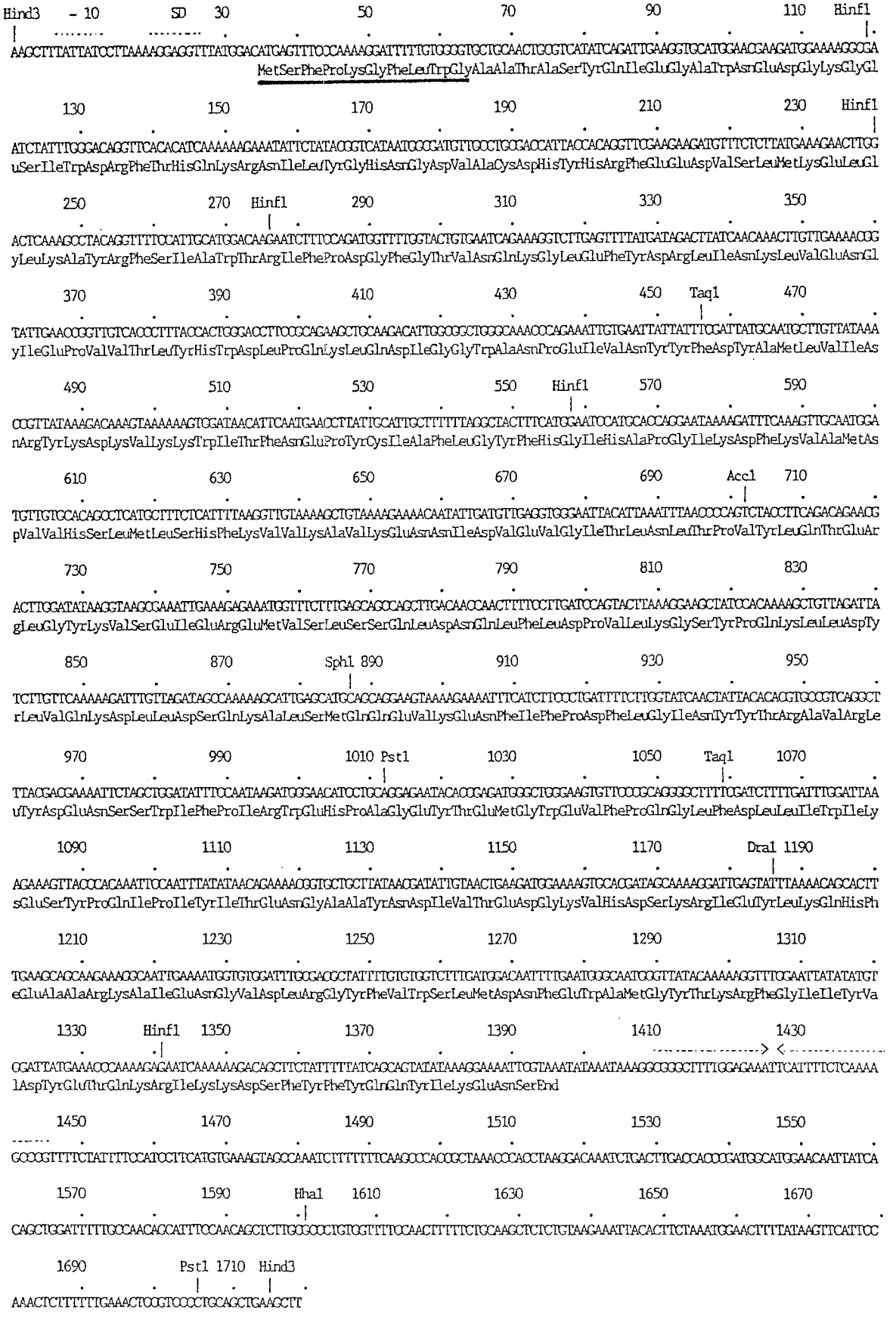

Fig. 1. Nucleotide and amino acid sequence of the *bglA* gene. Numbering of the nucleotide sequence begins at the *bglA* amino-terminus at the proximal end of the *Hin*dIII fragment. The *dots* mark every tenth nucleotide. Some restriction enzyme cut sites are indicated. The *underlined* amino acid sequence has been determined by automated sequencing of the purified protein. A putative Pribnow box (–10) and Shine-Dalgarno (S-D) sequence are indicated as *dotted lines*. A palindromic sequence is indicated by *dotted arrows* facing each other; this sequence does not qualify as a transcription terminator according to computer analysis using the Brendel-Trifonov algorithm (Brendel and Trifonov 1984)

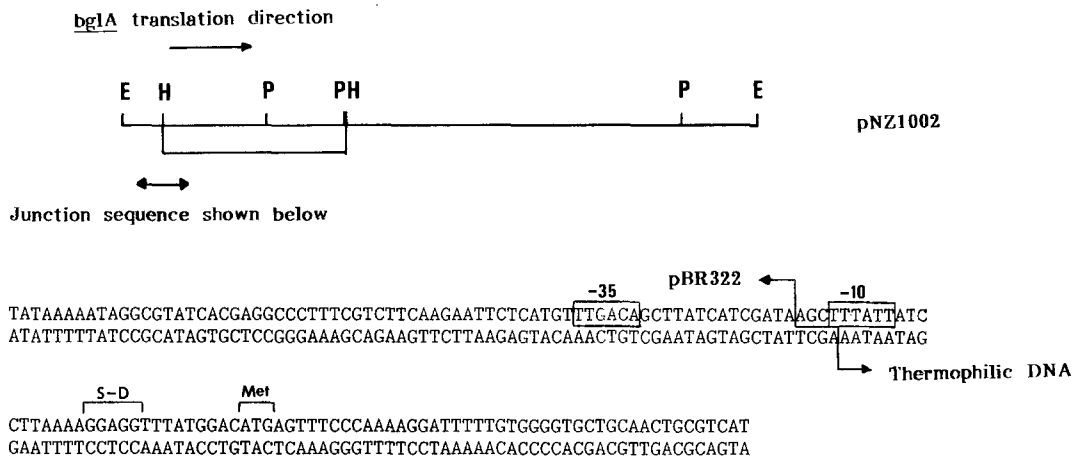

Fig. 2. Possible promoter and ribosomal binding site sequences formed by cloning the 1.72 kb *Hin*dIII fragment of pNZ1002 or pNZ1001, coding for β-glucosidase, into the *Hin*dIII site of pBR322. β-Glucosidase activity is expressed in both orientations but at very different levels (Love and Streiff 1987). The sequences of the joint-point of the construct encompassing the β-glucosidase ATG site are shown. The –35 and –10 sequences identified by the Targsearch program are *boxed*. *S-D* refers to the putative Shine-Dalgarno sequence and the translation start site (*Met*) is also indicated. Abbreviations are: E, *Eco*RI; P, *Pst*I; H, *Hin*dIII

SDS-molecular weight markers were purchased from Sigma.

Protein microsequencing. This was performed by Dr. D. Christie of the Department of Biochemistry, University of Auckland, using an Applied Biosystems Sequenator.

DNA sequence analysis. The 1.72 kb *Hin*dIII fragment containing the β-glucosidase gene was ligated into *Hin*dIII-digested RF (replicative form) of mp19 (Norrander et al. 1983) and recombinant DNAs were isolated with the thermophilic DNA fragment in both orientations. Single-stranded template DNAs were prepared from two isolates with opposite orientations of the *Hin*dIII fragment and deletions were prepared using T4 DNA polymerase (Dale et al. 1985). Individual deletion derivatives were sequenced using the dideoxy procedure (Sanger et al. 1977).

Computer analysis. All analysis of the sequence data was carried out using the Sequence Analysis Software Package of the University of Wisconsin Genetics Computer Group on a MicroVax II.

Results

The DNA sequence and the deduced polypeptide sequence are shown in Fig. 1. It can be seen that there are two ATG codons, one at positions 29–31 and one at 35–37. The amino-terminal sequence of the enzyme was determined on a purified sample of the protein isolated from *E. coli* cells transformed with is pNZ1001, to help determine which start codon is used (see Fig. 2). The sequence was found to be Ser-Phe-Pro-Lys-Gly-Phe-Leu-Trp-Gly-, with a small proportion of the protein giving the sequence Met-Ser-Phe-Pro-Lys, etc. It would appear that, at least in *E. coli*, there is amino-terminal processing of the β-glucosidase. It is likely that translation starts at the ATG codon at positions 35–37, since Ben-Bassat and Bauer (1987) have shown that amino-terminal methionine may be removed in vivo by methionine aminopeptidase particularly when the next residue is a serine. Inspection of Fig. 1 shows that there is no sequence of basic and hydrophobic amino acids characteristic of signal sequences. The open reading frame extending downstream of the second ATG codon could encode a polypeptide of 453 amino acids and with a molecular weight of 54400, which is close to the M_r of 52000 observed in experiments with maxicells (Love and Streiff 1987).

The orientation-dependent expression of the *Caldocellum* β-glucosidase in *E. coli* suggested that it was expressed from a vector promoter (Love and Streiff 1987). Fig. 2 shows that the *Hin*dIII fragment containing the *bglA* gene can be ligated into the *tet* promoter of pBR322 so that the –35 sequence is provided by the vector and the –10 sequence and ribosome binding site by the insert (Fig. 2). Analysis of the sequence so formed using the 'Targsearch' program (Mulligan et al. 1984) gave a promoter of moderate strength (score of 59.2%). A consensus Shine-Dalgarno sequence (AGGAGG for *E. coli*) is present in the *Caldocellum* DNA 9 bases upstream from the translational start site.

Sequence analysis of the *Caldocellum* DNA upstream of the *Hin*dIII junction showed that there is a potential –35 sequence separated by 17 bases from the –10 sequence shown in Fig. 2. This combined sequence scores as a moderately strong promoter using the 'Targsearch' program (58%). We concluded that *C. saccharolyticum* may use promoter sequences that are very similar to those used by the unrelated mesophile, *E. coli*.

The gene inserted in the opposite orientation is able to be expressed from the P1 ("anti-tet") promoter sequence of pBR322 (not shown).

In vivo transcription analysis of the β-glucosidase gene

Computer analysis of the DNA sequence of the 1.72 kb *Hin*dIII fragment indicated a putative –10 sequence immediately upstream of the translation start site of the *bglA* gene (Fig. 1). The vector pK100, a derivative of pKO-1 (McKenny et al. 1981), does not have a promoter upstream of its *Hin*dIII cloning site. The *Hin*dIII fragment was ligated in both orientations into this vector to determine whether the –10 sequence alone was biologically active, since tran-

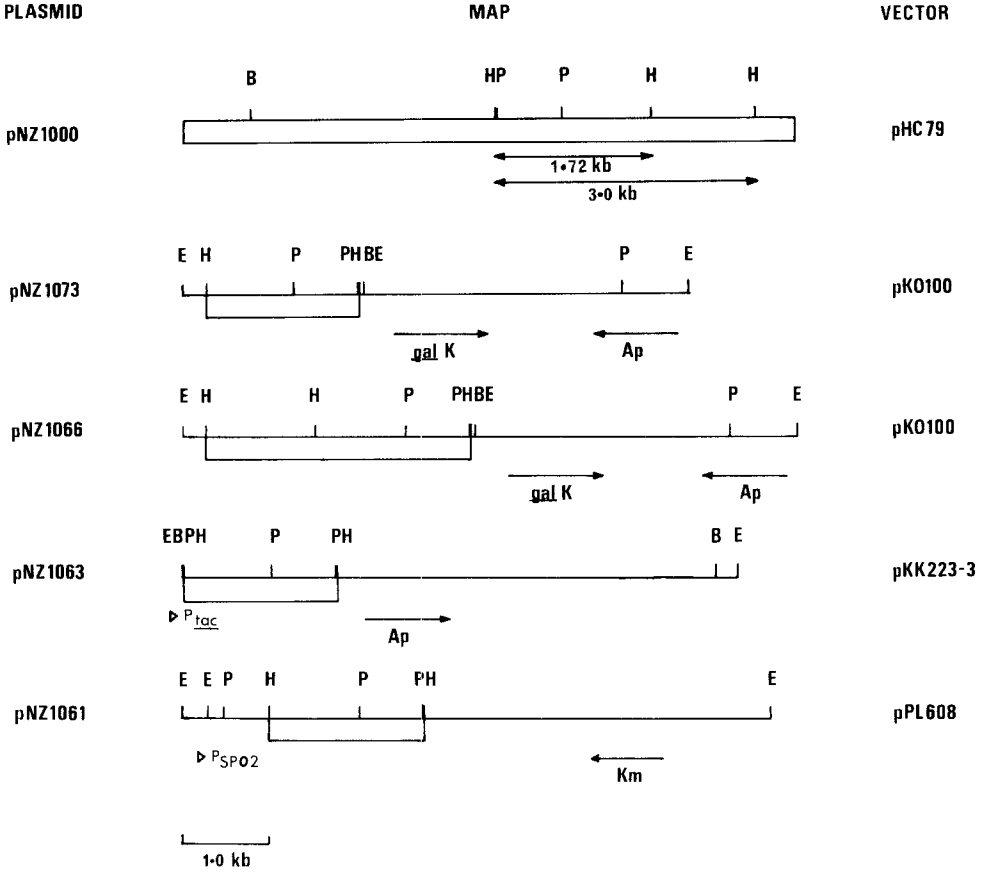

Fig. 3. Construction and restriction enzyme maps of recombinant plasmids containing the β-glucosidase gene. (i) pNZ1000 is a pHC79 recombinant containing a partial *Sau*3A fragment of *Caldocellum saccharolyticum* chromosomal DNA ligated into the *Bam*HI site of the cosmid vector pHC79 (Hohn and Collins 1980). Only the thermophilic DNA insert (7 kb) of pNZ1000 is shown. (ii) pNZ1000 was digested partially with *Hin*dIII and the 1.72 kb and 3 kb fragments were isolated and ligated into the *Hin*dIII site of pKO100. Plasmids were isolated from ampicillin-resistant *Escherichia coli* PB2636 transformants. Two kinds of recombinant plasmids containing the thermophilic DNA inserts isolated were named pNZ1066 and pNZ1073, and those ligated in the opposite orientation were called pNZ1065 and pNZ1074, respectively. (iii) The 1.732 kb *Hin*dIII fragment of pNZ1001 (Fig. 2) was ligated into the *Hin*dIII sites of pKK223-3 and pPL608. Plasmids were isolated from ampicillin (Ap)-resistant *E. coli* JM105 (for pKK223-3) and kanamycin (Km)-resistant, chloramphenicol-sensitive *Bacillus subtilis* SB202 (for pPL608) transformant colonies. The maps of pNZ1063 and pNZ1061 are shown. Recombinant plasmids containing the thermophilic DNA inserts ligated in the opposite orientation were called pNZ1064 and pNZ1062, respectively. The *tac* and SP02 promoters are shown and the direction of transcription from these promoters is indicated by an *open arrow head.* The arrows indicate the direction of transcription. *gal*K refers to the galactokinase gene. Abbreviations are: B, *Bam*HI; E, *Eco*RI; H, *Hin*dIII; P, *Pst*I

Table 1. Expression of β-glucosidase in *Escherichia coli* (PB2636) cells containing pKO100 recombinant plasmids. Transformed *E. coli* was grown overnight in the presence of ampicillin (50 μg ml^{-1}) at 37° C. Samples of the cultures were treated with toluene and assayed for β-glucosidase activity at 70° C using *p*-nitrophenyl-β-D-glucopyranoside (PNPG) as substrate

Plasmid	Units β-glucosidase activity (mg protein)$^{-1}$
pKO100	0
pNZ1073	0.025
pNZ1074	0.006
pNZ1065	2.299
pNZ1066	2.087

scription is dependent on the presence of a promoter sequence within the inserted DNA. The amount of β-glucosidase activity expressed by *E. coli* PB2636 cells transformed with either recombinant plasmid (pNZ1073 and pNZ1074, Fig. 3) was just above the background level (Table 1). This result indicates that the putative −10 sequence alone is insufficient to direct the expression of β-glucosidase in *E. coli.*

We had constructed earlier a pHC79 recombinant plasmid called pNZ1000 (Fig. 3), which contained the *bgl*A gene and expressed β-glucosidase activity in *E. coli* (data not shown). This plasmid was used to determine whether DNA upstream of the β-glucosidase gene contained the promoter sequence recognised in the mesophilic host. A 3 kb fragment from a partial *Hin*dIII digest of pNZ1000, which contained the 1.72 kb fragment and its upstream neighbour of 1.2 kb was isolated and ligated into pKO100 in both orientations. Plasmids pNZ1065 and and pNZ1066 expressed approximately equal levels of β-glucosidase activity which were significantly greater than the levels expressed by the plasmids containing only the 1.72 kb *Hin*dIII fragment (Table 1). These data confirmed the conclusion from sequence analysis that a promoter sequence which is recognised in *E. coli* lies upstream of the β-glucosidase gene and that at least the −35 sequence is present in the 1.2 kb *Hin*dIII fragment (discussed previously).

The direction of transcription of the *bglA* gene in

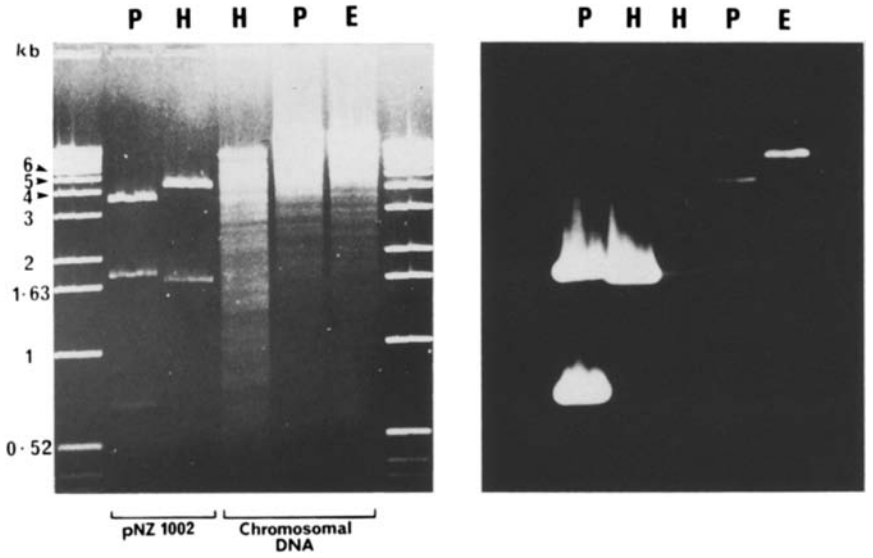

Fig. 4. A, B Southern blot hybridization of digested chromosomal DNA and plasmid pNZ1002 with the 1.72 kb *Hin*dIII fragment. Plasmid pNZ1002 and *C. saccharolyticum* chromosomal DNA were digested with several restriction enzymes and then electrophoresed in a 0.8% agarose-borate gel. The outside lanes contained a 1 kb DNA ladder (Bethesda Research Laboratories); the sizes of DNA markers are indicated at the left side. The DNAs were blotted to Genescreen Plus and hybridized with the 1.72 kb *Hin*dIII fragment of pNZ1002 which had been radioactively-labelled with [α-^{32}P] dCTP. **A** The agarose gel stained with ethidium bromide. **B** The radioautograph. Abbreviations are: E, *Eco*RI; H, *Hin*dIII; P, *Pst*I

pNZ1066 (see Fig. 3) would be expected to allow the expression of the promoterless *galK* gene of pKO100. However, plasmid pNZ1066 expressed no detectable galactokinase activity as determined by plating transformed *E. coli* PB2636 cells on MacConkey- galactose plates containing ampicillin. This lack of activity indicates that a sequence is present downstream of the β-glucosidase gene which prevents transcription proceeding into the *galK* gene. An inverted repeat sequence following the *bglA* structural gene was identified (Fig. 1), but as it was not similar to the rho factor independent terminator, its significance is unknown.

Hybridization analysis

Multiple genes coding for endocellulases have been isolated from several cellulolytic micro-organisms (Béguin et al. 1987; Romaniec et al. 1987a; Bergquist et al. 1987). In some cases the cloned cellulases showed significant homology although they were isolated on unique restriction enzyme fragments and are thus members of a single gene family. We used the 1.72 kb *Hin*dIII fragment containing the gene as a probe to establish whether or not there were multiple copies of the β-glucosidase gene present on the chromosome of *C. saccharolyticum*. Hybridization of the radioactively labelled probe showed that it hybridized with an 11 kb *Eco*RI fragment and two *Pst*I fragments (5 kb and 0.7 kb). The smaller *Pst*I fragment is the same size as an internal *Pst*I fragment within the probe DNA (Fig. 4). From these results we concluded that there is a single *bglA* gene (or a tandem repeat of it) in the genome of *C. saccharolyticum*.

Over-expression of β-glucosidase in mesophilic hosts

The over-production of thermophilic β-glucosidase in mesophilic hosts was attempted to enable the purification of large amounts of this enzyme. The 1.72 kb *Hin*dIII fragment was ligated in both orientations into the *Hin*dIII site of the expression vectors pKK223-3 (*E. coli*) and pPL608 (*B. subtilis*), (Fig. 3). The pKK223-3 vector contains the *tac* promoter (de Boer et al. 1983) and the expression of a gene inserted immediately down-stream of this promoter is regulated by the addition of isopropyl-β-D-thiogalactoside (IPTG). Plasmid pPL608 (Williams et al. 1981a) contains a strong phage promoter which functions in *B. subtilis*. The expression of a foreign gene inserted at the *Hin*dIII site of pPL608 is inducible by the addition of sub-inhibitory concentrations of chloramphenicol (Williams et al. 1981b).

Cells carrying pKK223-3, pPL608 and the recombinant plasmids containing the *bglA* gene were grown under antibiotic selection to mid-log phase. Each culture was divided and IPTG or chloramphenicol was added to one portion. Preliminary experiments had allowed the concentration of inducer necessary for maximum induction of β-glucosidase activity to be determined and had shown that the addition of inducer had no inhibitory effect on cell growth (data not shown).

Figure 5 shows that a 10-fold and 1.5-fold increase in the level of β-glucosidase activity expressed by pNZ1063 and pNZ1061, respectively, was detected 240 min after induction. However, the induced level of β-glucosidase activity in *B. subtilis* was greater than that in *E. coli*: 21.03 units mg^{-1} protein compared with 3.69 units mg^{-1}. The vectors pPL608 and pKK223-3 and the *E. coli* recombinant with the 1.72 kb *Hin*dIII fragment in the opposite orientation, pNZ1064, expressed no detectable β-glucosidase activity. The *B. subtilis* recombinant plasmid pNZ1062, with the *Hin*dIII fragment in the opposite orientation to pNZ1061, expressed 0.31 units ml^{-1} in the absence, and 0.25 units ml^{-1} after 240 min in the presence of chloramphenicol.

SDS-polyacrylamide gel electrophoresis of samples of induced and non-induced *B. subtilis* cells carrying pPL608, pNZ1061, and pNZ1062 confirmed the induction characteristics of β-glucosidase expression determined by enzyme assay. The arrow in Fig. 6 indicates a protein of M_r 51000

Table 2. Codon usage of *bglA* gene of *Caldocellum saccharolyticum*. The data for *E. coli* genes (highly expressed) comes from the University of Wisconsin Computer Group software package. The fraction for *Thermus thermophilus* isopropylmalate (IPM) dehydrogenase has been calculated from data in Oshima (1986)

Amino acid	Codon	Number of codons	Fraction	*Escherichia coli* fraction	*Thermus thermophilus* IPM dehydrogenase fraction
Gly	GGG	1	0.03	0.02	0.53
Gly	GGA	11	0.37	0	0.17
Gly	GGT	11	0.37	0.59	0
Gly	GGC	7	0.23	0.38	0.31
Glu	GAG	4	0.12	0.22	0.93
Glu	GAA	29	0.88	0.78	0.07
Asp	GAT	20	0.69	0.33	0
Asp	GAC	9	0.31	0.67	1.00
Val	GTG	9	0.30	0.16	0.63
Val	GTA	7	0.23	0.26	0
Val	GTT	11	0.37	0.51	0
Val	GTC	3	0.10	0.07	0.27
Ala	GCG	1	0.05	0.26	0.29
Ala	GCA	13	0.59	0.28	0.02
Ala	GCT	5	0.23	0.35	0.02
Ala	GCC	3	0.14	0.10	0.67
Arg	AGG	6	0.35	0	0.21
Arg	AGA	7	0.41	0	0
Ser	AGT	2	0.10	0.03	0
Ser	AGC	10	0.48	0.20	0.33
Lys	AAG	8	0.24	0.26	0.94
Lys	AAA	25	0.76	0.74	0.06
Asn	AAT	12	0.55	0.06	0
Asn	AAC	10	0.45	0.94	1.00
Met	ATG	10	1.00	1.00	1.00
Ile	ATA	8	0.24	0	0.11
Ile	ATT	19	0.58	0.17	0
Ile	ATC	6	6.18	0.83	0.89
Thr	ACG	0	0	0.07	0.54
Thr	ACA	8	0.53	0.04	0
Thr	ACT	3	0.20	0.35	0
Thr	ACC	4	0.27	0.55	0.46
Trp	TGG	13	1.00	1.00	1.00
End	TGA	0	0	0	0
Cys	TGT	0	0	0.49	0
Cys	TGC	2	1.00	0.51	0
End	TAG	0	0	0	0
End	TAA	1	1.00	0	1.00
Tyr	TAT	21	0.68	0.25	0.17
Tyr	TAC	10	0.32	0.75	0.83
Leu	TTG	7	0.18	0.03	0.08
Leu	TTA	6	0.16	0.02	0.03
Phe	TTT	15	0.56	0.24	0.25
Phe	TTC	12	0.44	0.76	0.75
Ser	TCG	1	0.05	0.04	0.07
Ser	TCA	1	0.05	0.02	0
Ser	TCT	6	0.29	0.34	0.07
Ser	TCC	1	0.05	0.37	0.53
Arg	CGG	0	0	0	0.36
Arg	CGA	2	0.12	0.01	0.07
Arg	CGT	2	0.12	0.74	0.04
Arg	CGC	0	0	0.25	0.32
Gln	CAG	11	0.58	0.86	1.00
Gln	CAA	8	0.42	0.14	0
His	CAT	8	0.58	0.17	0
His	CAC	5	0.42	0.83	1.00

Table 2 (continued)

Amino acid	Codon	Number of codons	Fraction	*Escherichia coli* fraction	*Thermus thermophilus* IPM dehydrogenase fraction
Leu	CTG	2	0.05	0.83	0.28
Leu	CTA	1	0.03	0	0.03
Leu	CTT	20	0.53	0.04	0.17
Leu	CTC	2	0.05	0.07	0.42
Pro	CCG	3	0.19	0.77	0.22
Pro	CCA	10	0.63	0.15	0
Pro	CCT	3	0.19	0.08	0.11
Pro	CCC	0	0	0	0.67
		453			

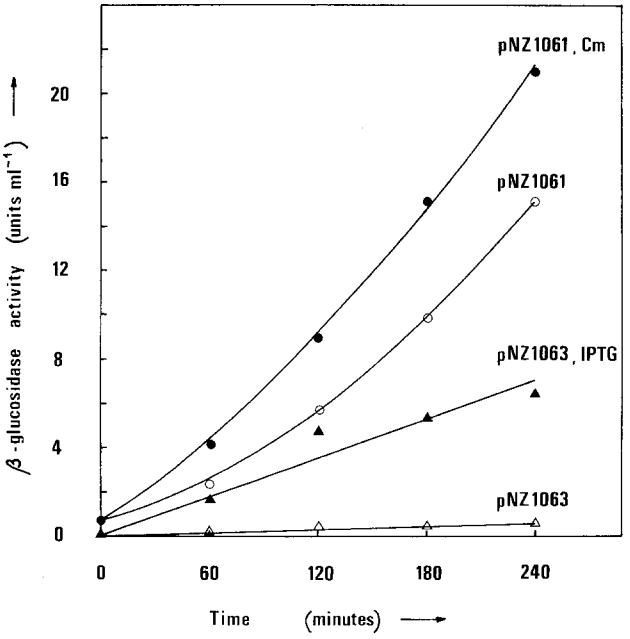

Fig. 5. Induction of β-glucosidase expression in *E. coli* JM105 [pNZ1063] and *B. subtilis* SB202 [pNZ1061]. *E. coli* and *B. subtilis* strains carrying *bglA* expression plasmids were grown to a density of $1–2 \times 10^8$ cells ml^{-1} in the presence of ampicillin (50 μg ml^{-1}) and kanamycin (5 μg ml^{-1}), respectively. The cultures were divided and IPTG (25 μm ml^{-1}) or chloramphenicol (0.1 μg ml^{-1}) was added to one of the *E. coli* and *B. subtilis* cultures, respectively. Samples were removed at hourly intervals and assayed for β-glucosidase activity

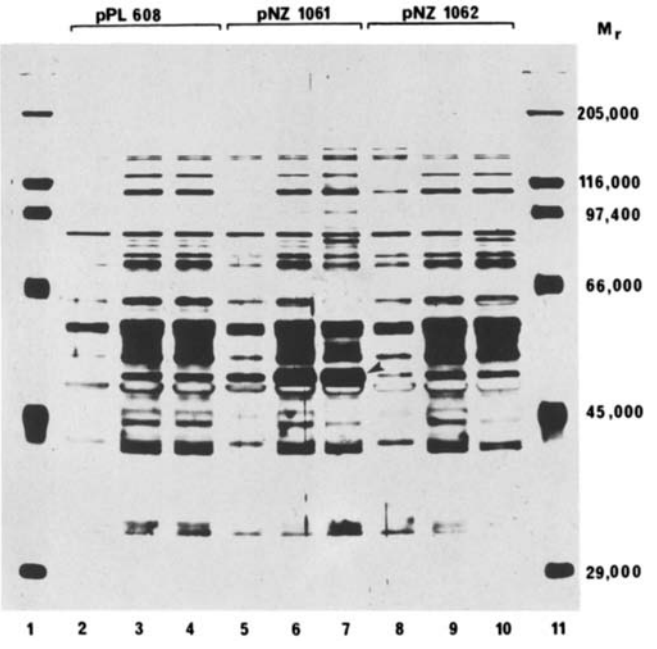

Fig. 6. SDS-polyacrylamide gel analysis of the induced expression of β-glucosidase in transformed *B. subtilis* SB202. Transformed *B. subtilis* SB202 was induced as described in Fig. 5. Samples were prepared and electrophoresed, and proteins were stained, as described in the Materials and methods. Lanes 2, 5 and 8: samples removed immediately prior to the addition of chloramphenicol (Cm); lanes 3, 6 and 9: samples removed after 240 min incubation in the absence of Cm; lanes 4, 7 and 10: the same as 3, 6 and 9 except that the cultures were incubated in the presence of Cm. Lanes 1 and 11 contain protein molecular weight standards (from top to bottom): myosin ($M_r = 205000$), β-galactosidase (M_r 116000), phosphorylase B (M_r 97400), bovine plasma albumin (M_r 66000), ovalbumin (M_r 45000) carbonic anhydrase (M_r 29000). The *arrow* indicates a protein of M_r 51000

which is expressed by pNZ1061 in the absence of chloramphenicol and is induced approximately 1.5-fold by the addition of chloramphenicol (lanes 6, 7). This protein is not expressed by pPL608 and pNZ1062 (Fig. 6, compare lanes 3 and 9 with lane 6). The apparent molecular weight of this protein is similar to that determined for β-glucosidase expressed by transformed *E. coli* maxicells (Love and Streiff 1987).

Discussion

Caldocellum saccharolyticum has a 34% G-C content and this fact is reflected in the sequence data and composition of the β-glucosidase gene (38% G-C). Consequently, some 4 bp recognition site enzymes cut rarely (for example, *Sau*3A, two cleavage sites) and there is no cleavage by enzymes with a high G-C content in their recognition sites. The β-glucosidase enzyme produced in *E. coli* is remarkably stable, with a half-life at 70° C of 2,280 min and a maximum assay temperature of 85° C. These values are significantly in excess of the temperature optima and stability of other β-glucosidases that have been examined (Bergquist et al. 1987). The temperature optimum of 85° C is 25° C higher than that shown by a β-glucosidase gene from *Clostridium thermocellum* cloned in *E. coli* (Romaniec et al. 1987b), and comparison of the sequences of the two enzymes may prove to be instructive for indicating residues contributing to ther-

mal stability. Unfortunately, no sequence data are available for the β-glucosidase from *Clostridium*.

Caldocellum saccharolyticum appears to be a Gram-positive organism, as shown by electron microscopy (W.H. Morgan, personal communication), and it resembles *Clostridium thermocellum* is this respect, as well as being another anaerobic thermophile. Genes involved in cellulose breakdown from both organisms share common features with Gram-positive and Gram-negative facultatively aerobic mesophiles in the structure of their transcribed DNA upstream of the protein initiation codon, for example, Pribnow box and strong Shine-Dalgarno sequences (Béguin et al. 1985; Joliff et al. 1986).

The sequences of the β-glucosidase genes from the fungi *Candida pelliculosa* and *Kluyveromyces lactis* have been reported recently (Kohchi and Toh-e 1987; Raynal et al. 1987). Computer analysis showed that there is no significant homology between these genes and the *bglA* gene of *Caldocellum*. Raynal et al. (1987) have commented on the similarity of the amino acid sequence of three fungal β-glucosidases in a peptide at the putative active site of each of the enzymes. Extensive computer comparisons of amino acid sequences showed no obvious similarity between the *Caldocellum* and the fungal β-glucosidases. Furthermore there is no homology with the *Clostridium thermocellum celA, celB* or *celD* genes and it does not contain the short re-iterated sequence possessed by these genes.

Table 2 shows the codon usage of the *bglA* gene. Oshima (1986) has reported that the G-C content of DNA of the extreme thermophile *Thermus thermophilus* is about 70%. The G-C content of the third letters of the codons for the 3-isopropylmalate dehydrogenase gene of *T. thermophilus* is about 90%. Oshima (1986) has pointed out that the optimal *E. coli* codons for valine, GUU and GUA, are not used in *T. thermophilus*. The information in Table 2 shows that codon usage in *Caldocellum* does not follow that of *Thermus* but resembles that of *E. coli*.

Sekiguchi et al. (1986) have compared the nucleotide and amino acid sequences of the 3-isopropylmalate dehydrogenase of *Saccharomyces cerevisiae* (mesophile), *Bacillus coagulans* (facultative thermophile) and *Thermus thermophilus* (extreme thermophile). They found that the G-C contents of the coding region and the third position of the codons of the *Bacillus* gene were intermediate in value compared to *S. cerevisiae* and *T. thermophilus*. The data for the β-glucosidase of *Caldocellum* does not fit any simple correlation between G-C content and thermostability (Oshima 1986).

It is now generally accepted that the enzymes of thermophiles are maintained in their native condition by intrinsic stability rather than by the presence of additional factors conferring thermal stability or by rapid turnover (Daniel 1986; Bergquist et al. 1987). There are no obvious indications from the deduced amino acid sequence of residues that particularly contribute to thermal stability, for example, amino acid substitutions like gly → ala within α-helical regions (Matthews et al. 1987). Oshima (1986) has suggested that thermostable proteins lack cysteine residues (as for example, *T. thermophilus* isopropylmalate dehydrogenase). However, *Caldocellum* β-glucosidase contains two cysteines and has a similar temperature optimum for enzymatic activity. Hence the presence of these amino acids does not seem to affect the thermostability of this protein.

Comparison of amino acid sequences and studies of the three-dimensional structures of a variety of proteins has shown that the greater heat stability of thermostable proteins is due to extra salt bridges between portions of the folded molecules (Perutz 1978; Daniel 1986). A conventional Chou-Fasman plot (Gribskov et al. 1986) of the β-glucosidase protein does not provide additional information as to which regions of the molecule are significant in thermostability. Our current experiments utilize segment-directed mutagenesis to generate mutations of several regions of the protein (Botstein and Shortle 1985; Matsumara et al. 1986).

Acknowledgements. This work was supported by grants from the Development Finance Corporation of New Zealand and the University of Auckland Research Committee. We are grateful for the capable assistance of Jan Robinson and Liam Williams.

References

Béguin P, Cornet P, Aubert J-P (1985) Sequence of a cellulase gene of the thermophilic bacterium *Clostridium thermocellum*. J Bacteriol 162:102–105

Béguin P, Millet J, Aubert JP (1987) The cloned *cel* (cellulase degradation) genes of *Clostridium thermocellum* and their products. Microbiol Sci 4:277–280

Ben-Bassat A, Bauer K (1987) Amino-terminal processing of proteins. Nature 326:315

Bergquist PL, Love DR, Croft JE, Streiff MB, Daniel RM, Morgan WH (1987) Molecular genetics and the biotechnological applications of thermophilic bacteria. Biotechnol Genet Eng Rev 5:199–244

Boer HA de, Comstock LJ, Vasser M (1983) The *tac* promoter: a functional hybrid derived from the *trp* and *lac* promoters. Proc Natl Acad Sci USA 80:21–25

Botstein D, Shortle D (1985) Strategies and applications of *in vitro* mutagenesis. Science 229:1193–1201

Brendel V, Trifonov EN (1984) A computer algorithm for testing potential prokaryotic terminators. Nucleic Acids Res 12:4411–4427

Dale RMK, McClure BA, Houchins JP (1985) A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing. Plasmid 13:31–40

Daniel RM (1986) The stability of proteins from extreme thermophiles. In: Oxender DI (ed) Protein structure, folding and design. Genex-UCLA Symposium, Liss, New York, pp 291–296

Donnison AM, Brocklesbury CM, Morgan WH (1986) A new species of non-sporulating thermophilic cellulolytic bacterium. Proc Int Congress Microbiol, Manchester, p 203

Gribskov M, Burgess RR, Devereux J (1986) PEPPLOT, a protein secondary structure analysis program for the UWGCG sequence analysis software package. Nucleic Acids Res 14:327–334

Hohn B, Collins J (1980) A small cosmid for the efficient cloning of large DNA fragments. Gene 11:291–298

Joliff G, Béguin P, Aubert J-P (1986) Nucleotide sequence of the cellulase gene *celD* encoding endoglucanase D of *Clostridium thermocellum*. Nucleic Acids Res 14:8605–8613

Kohchi C, Toh-e A (1987) Nucleotide sequence of *Candida pelliculosa* β-glucosidase gene. Nucleic Acids Res 13:6273–6282

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 277:680–685

Love DR, Streiff MB (1987) Molecular cloning of a β-glucosidase gene from an extremely thermophilic anaerobe in *E. coli* and *B. subtilis*. Biotechnology 5:384–387

Matsumara M, Kataoka S, Aiba S (1986) Single amino acid replacements affecting the thermostability of kanamycin nucleotidyltransferase. Mol Gen Genet 204:355–358

Matthews BW, Nicholson H, Bechtel WJ (1987) Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. Proc Natl Acad Sci USA 84:6663–6667

McKenney K, Shimitake H, Court D, Schmeissner U, Brady C, Rosenberg M (1981) A system to study promoter and terminator signals recognized by *Escherichia coli* RNA polymerase. In: Chirikjian J, Papas T (eds) Gene amplification and analysis, vol 2. Structural analysis of nucleic acids. Elsevier, New York, pp 383–415

Mulligan ME, Hawley DK, Entriken R, McClure WR (1984) *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. Nucleic Acids Res 12:789–800

Norrander J, Kempe T, Messing J (1983) Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. Gene 26:101–106

Oakley BR, Kirsch DR, Morris NR (1980) A simplified ultrasensitive silver stain for detecting proteins in polyacrylamide gels. Anal Biochem 105:361–363

Oshima T (1986) The genes and genetic apparatus of extreme thermophiles. In: Brock TD (ed) Thermophiles: general, molecular and applied microbiology. John Wiley and Sons, pp 137–147

Perutz M (1978) Electrostatic effects in proteins. Science 201:1187–1191

Raynal A, Gerbaud C, Francingues MC, Guerineau M (1987) Sequence and transcription of the β-glucosidase gene of *Kluyveromyces fragilis* cloned in *Saccharomyces cerevisiae*. Curr Genet 12:175–184

Reed KC, Mann DA (1985) Rapid transfer of DNA from agarose gels to nylon membranes. Nucleic Acids Res 13:7207–7221

Romaniec MPM, Davidson K, Hazlewood GP (1987a) Cloning and expression in *Escherichia coli* of *Clostridium thermocellum* DNA encoding β-glucosidase activity. Enzyme Microbiol Technol 9:474–478

Romaniec MPM, Clarke NG, Hazlewood GP (1987b) Molecular cloning of *Clostridium thermocellum* DNA and the expression of further novel endo-β-1,4-glucanase genes in *Escherichia coli*. J Gen Microbiol 133:1297–1307

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Sekiguchi T, Ortega-Cesena J, Nosoh Y, Ohashi S, Tsuda K, Kanaya S (1986) DNA and amino-acid sequences of 3-isopropylmalate dehydrogenase of *Bacillus coagulans*. Comparison with the enzymes of *Saccharomyces cerevisiae* and *Thermus thermophilus*. Biochim Biophys Acta 867:36–44

Williams DM, Duvall EJ, Lovett PS (1981a) Cloning restriction fragments that promote expression of a gene in *Bacillus subtilis*. J Bacteriol 146:1162–1165

Williams DM, Schoner RG, Duvall EJ, Preis LH, Lovett PS (1981b) Expression of *Escherichia coli trp* genes and the mouse dihydrofolate reductase gene cloned in *Bacillus subtilis*. Gene 16:199–206

Communicated by M. Takanami

Received January 7, 1988

Note added in proof

The *bgl*A gene inserted into pBR322 in the opposite orientation to that shown in Fig. 1 (pNZ1001) results in the −35 and −10 sequences being provided by the vector and the ribosome binding site by the thermophile DNA. The Targsearch programme scores this construct as a moderately strong promoter (58.6%, compared to 59.2% for pNZ1002). However, expression in *E. coli* of β-glucosidase by pNZ1001 is much greater than for pNZ1002. Sequence data suggests that the promoter structure in *Caldocellum* itself is virtually identical to that shown in Fig. 2.